Bioinformatics in bacterial molecular epidemiology and public health

Carrico, J. A.; Sabat, A. J.; Friedrich, A. W.; Ramirez, M.; ESGEM

Published in:
Eurosurveillance

Publication date:
2013

Bioinformatics in bacterial molecular epidemiology and public health: databases, tools and the next-generation sequencing revolution

J A Carriço (jcarrico@fm.ul.pt)[1], A J Sabat[2], A W Friedrich[2], M Ramirez[1], on behalf of the ESCMID Study Group for Epidemiological Markers (ESGEM)[3]

1. Instituto de Microbiologia, Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa, Lisbon, Portugal
2. Department of Medical Microbiology, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands
3. European Society for Clinical Microbiology and Infectious Diseases, Basel, Switzerland

Citation style for this article:
Carriço JA, Sabat AJ, Friedrich AW, Ramirez M, on behalf of the ESCMID Study Group for Epidemiological Markers (ESGEM). Bioinformatics in bacterial molecular epidemiology and public health: databases, tools and the next-generation sequencing revolution. Euro Surveill. 2013;18(4):pii=20382. Available online: http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=20382

Article submitted on 29 June 2012 / published on 24 January 2013

Advances in typing methodologies have been the driving force in the field of molecular epidemiology of pathogens. The development of molecular methodologies, and more recently of DNA sequencing methods to complement and improve phenotypic identification methods, was accompanied by the generation of large amounts of data and the need to develop ways of storing and analysing them. Simultaneously, advances in computing allowed the development of specialised algorithms for image analysis, data sharing and integration, and for mining the ever larger amounts of accumulated data. In this review, we will discuss how bioinformatics accompanied the changes in bacterial molecular epidemiology. We will discuss the benefits for public health of specialised online typing databases and algorithms allowing for real-time data analysis and visualisation. The impact of the new and disruptive next-generation sequencing methodologies will be evaluated, and we will look ahead into these novel challenges.

Introduction

In the past twenty years, the advances in several fields of biology, molecular biology in particular, led to an increased capacity to generate data. This resulted in the accumulation of large datasets and the need to store, manage and analyse them. This was the starting point for the development of the multidisciplinary field of bioinformatics. Hesper and Hogeweg originally coined the term bioinformatics in 1970 [1]. It was broadly defined as "the study of informatics processes in biotic systems". But it was the convergence of mathematicians, computer scientists, physicists, biologists, chemists and health professionals for the analysis of the biological data generated in the genomic revolution that resulted in the diverse disciplines comprised within bioinformatics. The field can also be subdivided into two large, interrelated subareas: data management, encompassing the creation and management of databases for biological data, and data analysis, ranging from the creation of mathematical and statistical models to computational tools and data mining techniques.

In bacterial molecular epidemiology, bioinformatics drove the creation of online databases for microbial typing data (e.g. antibiotic resistance profiles, phage typing, serotyping or other phenotypic information), the analytic methodologies for gel-based molecular typing techniques and the study and analysis of phylogenetic inference models.

In this review we aim to provide a perspective on the bioinformatics tools that have been applied in the field of bacterial molecular epidemiology. We will explore their applications in public health, documenting how they have changed and discussing possible avenues for future research and development in the field.

Online databases for bacterial typing

Microbial typing methods allow the characterisation of bacteria to the strain level, providing researchers with important information for surveillance of infectious diseases, outbreak investigation and control. These methods offer insights into the pathogenesis and natural history of an infection, and into bacterial population genetics [2,3], areas of research that have an important impact on human health issues such as the development of vaccines or novel antimicrobial drugs [4], with significant social and economical implications.

Molecular typing methods, such as pulsed-field gel electrophoresis (PFGE), provided the intra- and inter-laboratory reproducibility needed to create databases of isolates that could be used for longitudinal studies [3]. This allowed for bacterial typing to extend beyond outbreak investigation. Results were originally stored in local databases, using specialised software

Table 1

Online molecular typing databases

Method	Database	URL
MLST	MLST.net	http://www.mlst.net
	Pubmlst.org	http://www.pubmlst.org
	Institut Pasteur MLST	http://www.pasteur.fr/mlst/
	European Working Group for Legionella Infections Sequence-based typing database	http://www.hpa-bioinformatics.org.uk/legionella/legionella_sbt/php/sbt_homepage.php
	Environmental Research Institute, University College Cork	http://mlst.ucc.ie/
MLVA	MLVAbank	http://minisatellites.u-psud.fr/MLVAnet/
	Groupe d'Etudes en Biologie Prospective	http://www.mlva.eu
	MLVAplus	http://www.mlvaplus.net/
	Institute Pasteur MLVA: MLVA-NET	http://www.pasteur.fr/mlva
	MLVA.net	http://www.mlva.net
ccrB typing	Staphylococci ccrB sequence typing	http://www.ccrbtyping.net/
dru typing	dru typing database	http://www.dru-typing.org
spa typing	Ridom Spa Server	http://spaserver.ridom.de/
CRISPR typing	CRISPRdb	http://crispr.u-psud.fr/crispr/

CRISPR: Clustered Regularly Interspaced Short Palindromic Repeat; MLST: multilocus sequence typing; MLVA: multilocus variable-number tandem repeat analysis.

such as BioImage Whole Band Analyzer (Genomic Solutions, Inc, Ann Arbor, MI, currently discontinued) and GelCompar (currently GelComparII or Bionumerics from Applied Maths, Ghent, Belgium). These pieces of software, which integrated rudimentary database management and gel image analysis, were in fact the first widely adopted bioinformatics tools used in the field.

The ability to share information using the Internet led to the next step: the evolution of those software applications to distributed systems in which nationwide or worldwide comparisons could be performed. PulseNet, the molecular subtyping network for foodborne bacterial disease [5] was the first network that created local and central databases where laboratories from across the United States (US), could securely query nationwide data and compare their local samples. PulseNet is a governmental network initiated by the US Centers for Disease Control and Prevention and laboratories in several state health departments in the US, but has evolved to PulseNet International (www.pulsenetinternational.org/) [6]. PulseNet was created based on standardised PFGE protocols for the identification of pathogenic food-borne bacteria, relying on specifically trained technical personnel, but nowadays also integrates information obtained by other typing methods.

The network derives its strength from a series of bioinformatics techniques, implemented in the Bionumerics software, that range from optimised algorithms for gel image analysis and comparison to database management and secure sharing of data. The PulseNet online information system became the first distributed database for microbial typing with a direct application in public health and remains an example of the successful application of bioinformatics in typing and molecular epidemiology.

What the PulseNet distributed network achieved for PFGE, was much more simply achieved for multilocus sequence typing (MLST) [7], due to the inherent portability of sequence data (i.e. data easily transferrable between different systems). MLST is based on the analysis of allelic profiles generated by comparing sequences to an online repository. In contrast to PulseNet, MLST websites host publicly accessible databases where any laboratory can submit data, while PulseNet is only accessible by their member laboratories due to privacy and confidentiality issues (Table 1).

The ability to easily share sequence data through the Internet [8,9] is one of the main characteristics that made MLST the method of choice for clonal identification and tracking for many bacterial species. Currently available MLST databases (Table 1) are more commonly used for nomenclature purposes and may not reflect clonal abundance. The portability that is characteristic of MLST allows disambiguation when analysing and comparing results. Another important feature that contributed to its success was the possibility to infer patterns of phylogenetic descent through comparison of the allelic profiles. Even though MLST became the gold standard for long-term epidemiological surveillance of several species, PFGE remains important for outbreak

detection because it often has higher discriminatory power.

One example of an MLST online database, with proven use in public health, is the European Working Group for Legionella Infections (EWGLI) database, currently part of the European Legionnaire's Disease Surveillance Network (ELDSnet). This typing scheme and database successfully identified sources of infection, by determining clonal identity between environmental and patient isolates of *Legionella pneumophila* [10].

Several other sequence-based typing methodologies with online databases have become available. In contrast to MLST, the majority of these methods are only available for certain species, since they focus on non-housekeeping genes, and most are single locus sequence typing (SLST) schemes.

Taking *Staphylococcus aureus* as an example, several SLST were developed in the past decade. Two methods based on variable-number of tandem repeats (VNTR) were proposed, one relying on the direct repeat unit (*dru*) VNTR region adjacent to IS*431* in SCC*mec* [11], and the other based on the analysis of repeat patterns in the *spa* gene, the now widely used *spa* typing [12]. A major factor for the widespread use of *spa* typing was the implementation of a user-friendly software, Ridom StaphType. This tool allows the automatic assignment of a *spa* type from a DNA sequence in Fasta format or directly from chromatograms, through comparison with the centralised SpaServer [13]. Another SLST is *ccrB* typing [14], originally developed for meticillin-resistant *S. aureus* (MRSA), but extended and applicable to all staphylococci containing the *mecA* gene, the determinant of meticillin resistance. Also this method benefits from online databases and tools.

A multilocus methodology that has recently shown promise for several bacterial species is multilocus VNTR analysis (MLVA). Similarly to MLST it produces a numeric profile, in this case of the number of repeats at each locus that can unambiguously identify a given strain (MLVA type). Its appeal derives from providing a highly discriminatory method that shows high congruence with MLST results for several bacterial species [15], but is less expensive since sequencing of the loci is not necessary. Databases for a variety of schemes and bacterial species have been made available by several institutions (Table 1). Some of these online databases offer users the possibility to create their own private or public database like MLVAbank [16], MLVAplus or MLVA-NET [17]. A particular application of an MLVA scheme is the MIRU-VNTRplus Internet application for *Mycobacterium tuberculosis* [18,19].

Recently, a new sequence-based typing methodology was proposed using clustered regularly interspaced short palindromic repeats (CRISPR), a specific family of DNA repeats, conferring resistance to foreign DNA such as plasmids and phages. A database and tools are also available online (Table 1).

With next generation sequencing (NGS) technologies comes the ability to quickly obtain complete or nearly complete genome sequences of thousands of individual strains. In spite of the great promise of these approaches, it is still unclear how whole-genome data on bacterial pathogens will be shared and used for bacterial population surveillance and possible applications in public health.

BIGSdb, is a database system recently proposed to handle NGS data of microbial genomes and perform analyses focused on extended MLST typing approaches, which can comprise thousands of genes, and also on other population analysis methodologies [20]. One such scheme is ribosomal MLST [21] that, by focusing on the same ribosomal genes, allows a universal characterisation of bacteria, encompassing all levels of bacterial diversity, from domain to strain.

In highly monomorphic and slowly evolving bacterial species such as *M. tuberculosis* or Bacillus anthracis, identification of single nucleotide polymorphisms (SNPs) by comparison to a defined archetypal strain, could also be a basis for analysis, imposing different requirements on an online database.

Tools for data analysis

The cornerstone of molecular epidemiology is the ability to compare the classification results obtained by a given typing method for two or more distinct isolates and to measure their relatedness. With that information, one can then support an epidemiological investigation or raise a hypothesis about phylogenetic relationship. In this section we will describe several of the techniques developed in the last decades and used in the analysis of molecular typing data.

The first methodologies used in analysis of the phenotypic and genotypic data, were classical techniques used in numerical taxonomy [22], a field pioneered by P. Sneath and R. Sokal. The most popular are hierarchical clustering methods, which result in a unique tree representing the relationships between isolates, commonly called dendrogram or phenogram. From that tree, groups of related isolates are defined by a similarity level cut-off. These are mathematical methods that were implemented in generic statistical software or custom-made computer programmes. However, for the analysis of gel-based typing data, an integrated solution of image analysis and normalisation was needed prior to data analysis. This led to the development of the first tools specific for the analysis of gel-based typing methods. They allowed the quantitative analysis of large numbers of isolates and their comparison with databases of already characterised strains for gel-based methodologies such as PFGE, random amplification of polymorphic DNA (RAPD) [23], amplified fragment length polymorphism (AFLP) [24] or any

Table 2

Currently available software for the analysis of typing results

Application	Software	URL	Availability
Gel analysis	GelCompare II	http://www.applied-maths.com/gelcompar-ii	Commercial
	Phoretix 1D	http://www.totallab.com/products/1d/	
	Gel-Pro Analyzer 4.5	http://www.mediacy.com/index.aspx?page=GelPro	
Sequence assembly and analysis	Lasergene	http://dnastar.com	
	CLCbio workbench	http://www.clcbio.com/products/clc-main-workbench/	
	Geneious	http://www.geneious.com/	
Multiple	Bionumerics	http://www.applied-maths.com/bionumerics	
Phylogenetic inference	eBURST v3	http://eburst.mlst.net	Freeware
	MEGA 5	http://megasoftware.net/	
	PHYLOViZ 1.0	http://www.phyloviz.net	
	Structure 2.3.3	http://pritch.bsd.uchicago.edu/structure.html	
	BAPS 5.4	http://www.helsinki.fi/bsg/software/BAPS/	
	ClonalFrame 1.2	http://www.xavierdidelot.xtreemhost.com/clonalframe.htm	
Typing methods comparison	Ridom Epicompar	http://www.ridom.de/epicompare/	
	Comparing Partitions	http://www.comparingpartitions.info	
Recombination assessment	RDP3	http://darwin.uvigo.es/rdp/rdp.html	
Sequence comparison and analysis	Mauve	http://gel.ahabs.wisc.edu/mauve	

restriction fragment length polymorphism (RFLP) methodology. Presently, the most widely used and complete software solution for the analysis of gel-based typing methods is the commercially available Bionumerics, as it incorporates several hierarchical clustering algorithms for the analysis of typing data, as well as algorithms for the analysis of DNA sequences (Table 2).

With the appearance of MLST, new analysis methodologies were developed that tried to incorporate a model of bacterial evolution and spread. eBURST (based upon related sequence types) [25] implements a simple model for the emergence of clonal complexes [26,27]: a given genotype increases in frequency in the population and becomes a founder clone, and this increase is accompanied by a gradual diversification of that genotype, by mutation or recombination, forming a cluster of phylogenetically related strains. Software that performs eBURST analysis is available as freeware (Table 2).

The eBURST algorithm was further extended by goeBURST [28], a global optimal implementation of the eBURST algorithm that guarantees a unique solution for the BURST rules, while simultaneously allowing an assessment of the validity of each drawn link. The goeBURST algorithm is not exclusive for the analysis of MLST sequence types (ST) and can also be used in the analysis of any other sequence-based typing method that produces an allelic profile, such as MLVA or even SNP data from NGS methods. goeBURST also clarified the relationship between BURST rules and the use of minimum spanning trees (MSTs), another commonly used method in the analysis of sequence-based typing methods. It showed that the definition of clonal complexes by goeBURST is identical to pruning an MST at a chosen number of differences in the profiles that are being compared. That MSTs are easy to interpret has made them one of the preferred representation methods of the relationships inferred from SNP data in a variety of studies [29-31]

Although eBURST or goeBURST have been used extensively and successfully for determining the genetic population structure of many bacterial species, they also have limitations. As with other methods of phylogenetic reconstruction, the BURST rules do not specifically take into account recombination. Recombination is increasingly recognised as a major force in bacterial evolution, and when it involves segments of DNA larger than the internal gene fragments analysed by MLST, this will lead to the presence of the same alleles in strains from different genetic lineages. Horizontal gene transfer can therefore result in STs that have similar allelic profiles due to recombination, rather than recent shared ancestry. This is particularly true for some bacterial species such as Enterococcus faecium and Burkholderia pseudomallei in which recombination occurs with very high rates [32]. In other instances, recombination was even shown to occur between different species of the same genus [33]. To highlight recombination occurring within the analysed fragments different methods can be used, many are implemented in the software RDP3 [34], while traditional phylogenetic methods are helpful in detecting recombination between different species. An important set of tools are implemented in the software MEGA (Molecular Evolutionary Genetics Analysis) [35].

For the analysis of *spa* typing data, an algorithm was proposed to create clonal complexes from the sequence of repeats, based on an evolutionary model of repeated excision and duplication as well as single nucleotide substitutions and indels (insertions or deletions) (EDSI) [36]. This approach is available in the BURP (based upon repeat pattern) algorithm [37], implemented in the Ridom StaphType software, but could also be applied to other VNTR analysis.

An important aspect in the analysis of typing data is the integration of the algorithm results with epidemiological data. This is usually done by annotation of the resulting trees or dendrograms. Bionumerics offers that possibility in its multiple analysis algorithms. The freely available PHYLOViZ software [38] offers a more dynamic interface for the integration of this information into a goeBURST analysis, in the expansion of the goeBURST rules to any number of loci and in MSTs.

In epidemiological studies, the spatial component is of great importance. The ability to monitor the geographic spread of clones at different levels (cities, countries, continents or worldwide) can provide a perspective of the dissemination of successful clones. The website www.spatialepidemiology.net provides users with a map-based interface that allows the display and analysis of epidemiological data for infectious diseases. It was used by the European Antimicrobial Resistance Surveillance System (EARSS) [39] to provide a genetic snapshot of the *S. aureus* population causing invasive disease in Europe, plotting *spa* typing data, antibiotic resistance and other epidemiologically relevant data [40]. The website can also be connected to the EpiCollect system [41], allowing the real-time collection and annotation of data using any browser or smartphone.

The growing availability of sequence data also led to the increased popularity of model-based statistical analysis approaches. These focus on the use of Bayesian theory to infer the most probable population structure. The software applications STRUCTURE [42], Clonalframe [43] and Bayesian Analysis of Population Structure (BAPS) [44,45] are freely available, but have high computational requirements for large datasets. STRUCTURE and BAPS were initially proposed for classical population genetic analysis and try to infer possible population structures by identifying admixture events in the population history. Clonalframe was proposed for the analysis of MLST sequence data or alignments of multiple bacterial genomes and takes into account the possibility of recombination between sequences. More recently, BAPS was also adapted to detect and represent recombination between different populations and subpopulations [46] using MLST sequences as input. These methodologies can provide a much finer picture of how the phenomena shaping population structure interact and how they influence the final population [47-49], but the computational needs and complex analysis of results still limit their application in the field of bacterial epidemiology.

Not all bioinformatics tools in molecular epidemiology were initially designed for clonal inference from typing data. Two freely available tools were developed with the goal of providing a quantitative comparison of typing methods. Ridom Epicompare is a stand-alone software that allows the calculation of Simpson's index of diversity [50] and 95% confidence intervals [51] for a typing method, and the concordance indexes of Rand [52], adjusted Rand [53] and Wallace [54] for the assessment of congruence between typing methods [55]. The website www.comparingpartitions.info extends the features of Epicompare, by implementing confidence intervals for Wallace [56] and adjusted Rand [57] indices, as well as an adjusted Wallace coefficient and respective 95% confidence intervals [58]. These discriminatory and concordance indexes are now being used for evaluating the adequacy of a method for epidemiological typing. More recently these indexes were used to evaluate cut-off criteria for defining groups. This was done for multilocus variable-number tandem repeat fingerprinting (MLVF) patterns for *S. aureus* typing, including analyses of outbreaks and strain transmission events [59] as well as for PFGE [60], and also for defining clones in *Staphylococcus epidermidis* [61].

Bioinformatics for molecular epidemiology: the way forward

The advances in the last two decades in DNA sequencing capacity and bioinformatics led to an increase in the number of databases and software tools for microbial typing methods. The ability to freely share sequence data over the Internet, pioneered by MLST databases, was the turning point for the definition of a common language for the identification of bacterial clones.

However, the currently available databases suffer from several drawbacks. In some cases, data submission and curation protocols still rely heavily on human input with the exchange of files by email or other non-automated processes that are prone to human error and lead to extended response times by curators. Another missing feature is the absence of application programming interfaces for automatic querying and of standardised data sharing formats. These limitations make data collation a difficult and laborious manual process that requires integrating data from different databases and preparing them for analysis by available software. Consequently, a wealth of data is left largely inaccessible and unexplored.

The first step in tackling these problems is the definition of a common language to exchange data between databases and between databases and software. This is the starting point for the creation of database interoperability, i.e. the ability of tools in one database to query another, allowing for transparent data integration.

Current concepts and technologies for data integration are focused in the Semantic Web [62] and Linked Open Data concepts [63]. These concepts envision a data-centric approach with loosely standardised formats for information exchange, based on explicit data descriptions [64]. To achieve these goals, an ontology of terms in the field must be explicitly described. Ontologies provide a formal, standardised representation of the data and the relationships between the data entities [65]. Recently, the prototype of an ontology for microbial typing was proposed and made publicly available at www.phyloviz.net/typon/ [66]. The use of the ontology and the concepts of Linked Data for the construction of webservices for data exchange and validation could prove fundamental for the integration of the present techniques with the new NGS methods. This would allow NGS databases and data analysis algorithms to be validated against the large body of data available in existing databases.

The potential of NGS technology to become the ultimate methodology for bacterial identification and typing has been recognised by the scientific community, and the first steps towards its application have been taken.

NGS data result from a plethora of different technologies, each with its own strengths and caveats [67]. Running a single NGS analysis of an isolate will generate an amount of data that is orders of magnitude greater than that generated by other typing methods. As an example, the reads of a single bacterial genome with 100-fold coverage, will occupy around 200 MB of disk space. To handle this amount of data requires a complex IT infrastructure that was not necessary before. This also generates computational challenges that must be addressed by specialised software. Cloud computing and the use of high performance computing facilities will mitigate this problem, but are not a substitute for optimised algorithms. Stimulating collaborations between computer scientists and mathematicians with interest in biological problems, and developing specific training programmes will be key to attaining this goal.

Since the technology has been in constant evolution and the algorithms are evolving with it, there is currently no stable pipeline for the analysis of NGS data [68]. Due to limited availability of expertise in this area, centralised hubs for NGS application in diagnosis and public health have been proposed [69]. As the technology matures, the situation may change, allowing the deployment of NGS at hospital level. Recent releases of commercial Windows-based software with a menu-driven approach are a first step towards this goal (Table 2). However, it is important to note that at the current pace of innovation in NGS, these platforms frequently incorporate already superseded versions of algorithms that are under constant development in UNIX-based counterparts, less user-friendly, but freely available.

There are already several successful applications of NGS to a variety of public health problems, ranging from outbreak or short-term epidemiology investigations, to the discovery of unsuspected zoonosis cases and long-term epidemiology studies.

An event that received considerable media coverage was the outbreak of Escherichia coli O:104 haemolytic-uraemic syndrome in Germany that started in May 2011. Due to the pioneering crowdsourcing efforts in annotating an early released genome of an outbreak isolate and subsequent follow-up analyses [70,71], it was possible to promptly develop a diagnostic PCR to identify outbreak isolates. Subsequent studies were able to propose that the outbreak strain, E. coli O104:H4, had emerged due to horizontal gene exchange, shedding novel light on the emergence of new pathogens [72].

A recent pilot study focusing on the nosocomial pathogens MRSA and Clostridium difficile evaluated the feasibility of using benchtop sequencers for outbreak detection and surveillance at hospital level [73]. The ability to further discriminate isolates grouped together by other typing methods allowed a better understanding of the chains of transmission and supported infection control measures. Similar results were achieved when tracing an MRSA outbreak in a neonatal ward [74].

Long-term epidemiological studies have also benefited from NGS technology. The evolution of extremely successful and clones with worldwide dissemination has been followed for MRSA and *Streptococcus pneumoniae* [75,76]. Using SNP to identify phylogenetic relationships, these studies mapped the acquisition of mobile genetic elements and the fast-paced evolution of surface antigens that had frequently confounded previous analyses.

Most intriguing was the use of NGS to identify a probable zoonotic origin for autochthonous leprosy cases in the southern United States [77]. The study identified a unique genotype in this geographic area that also occurred in the armadillo population, strongly suggesting a zoonotic origin and a potential avenue for the control of this infection.

Two recent international meetings discussed and defined roadmaps in bacterial genomic identification and outbreak detection through the use of NGS.

The National Food Institute at the Technical University of Denmark issued a consensus report from an expert meeting on the perspectives of a global, real-time microbiological genomic identification system [78]. In this report it is recognised that within 5 to 10 years, DNA sequencers will likely be a common tool in clinical microbiology laboratories, and that the limiting factor will not be the cost of whole genome sequencing, but the creation of standardised pipelines to handle

the large amounts of data generated. It was also highlighted that a clear and widely accepted concept of the term 'clone' was needed, and that the comparison with data from existing databases (for example MLST) will play a crucial part in validating whole-genome approaches and providing the link with currently accepted and validated methodologies. It was further recognised that achieving this goal required "a global system or at least inter-operable systems to aggregate, share, mine and translate the genomic data to direct part of the genomics efforts to address global public health and clinical challenges, a high impact area in need of focused effort" [78].

A follow-up meeting was held in Washington, under the auspices of the United States Food and Drug Administration, also with the objective of establishing consensus guidelines in the field, focusing on NGS technology for outbreak detection. One of the most debated topics was the future development of databases for NGS data. The need for publicly available data repositories with NGS data from all bacterial domains was reinforced as a prerequisite for the development of new analysis methods.

These needs were also recently recognised in an expert consultation on molecular epidemiology hosted and organised by the European Centre for Disease Prevention and Control (ECDC) [79].

As more data becomes available, it is clear that molecular epidemiology will also benefit from closer integration with basic research in evolution and population biology. Changes in databases and analysis tools will be needed to bring about this integration in order to empower stakeholders in everyday public health decisions.

Tools are being developed to integrate different sources of molecular epidemiology data as well as other metadata (place, time, etc). However, these efforts are still in their infancy, and greater emphasis will need to be placed on the integration of different information sources in the analysis algorithms. Through the combined analysis of this information we can obtain knowledge of the epidemiology of infectious diseases. In particular, the broader use of geographic information in phylo-geographical approaches will allow a better understanding of the spread of particular clones [80].

Conclusions

Epidemiology has come a long way since John Snow investigated the cholera epidemic in Soho, London, in 1854. From hand-plotting cases on a map, we have come to depend on computing power and complex bioinformatics algorithms to make sense of the wealth of available molecular epidemiology data. It is clear that bioinformatics tools have raised the public health impact of the widely used typing methods. Similarly, the NGS revolution will not be extensively available to health professionals until several bioinformatics challenges have been solved and the results can be reported in a way that can be acted upon in everyday practice.

Integration of data of already established microbial typing methods, genomic and epidemiological databases and NGS data will be the next frontier in bacterial epidemiology. Once NGS becomes widely adopted, the development of software that analyses information from different data sources will be key to the synthesis of available knowledge. The public health community must also define standards for analysis and reporting, in order to produce the desired reproducibility and common language needed for typing based on NGS to be useful in clinical settings. More than ever, the need for a convergence of specialists of numerous disciplines in the field of bioinformatics will be fundamental to solve these challenges.

References

1. Hesper B, Hogeweg P. Bioinformatica: een werkconcept [Bioinformatics: a working concept]. Kameleon 1970; 1(6):28–9. Dutch.
2. van Belkum A, Struelens M, de Visser A, Verbrugh H, Tibayrenc M. 2001. Role of genomic typing in taxonomy, evolutionary genetics, and microbial epidemiology. Clin Microbiol Rev. 2001;14(3):547–60.
3. Struelens M. 1996. Consensus guidelines for appropriate use and evaluation of microbial epidemiologic typing systems. Clin Microbiol Infect 1996;2(1):2-11.
4. McKnew DL, Lynn F, Zenilman JM, Bash MC. Porin variation among clinical isolates of Neisseria gonorrhoeae over a 10-year period, as determined by Por variable region typing. J Infect Dis. 2003;187(8):1213–22.
5. Swaminathan B, Barrett TJ, Hunter SB, Tauxe RV, CDC PulseNet Task Force. PulseNet: the molecular subtyping network for foodborne bacterial disease surveillance, United States. Emerg Infect Dis 2001;7(3):382–9.
6. Swaminathan B, Gerner-Smidt P, Ng L-K, Lukinmaa S, Kam K-M, Rolando S, Gutiérrez EP, Binsztein N. Building PulseNet International: an interconnected system of laboratory networks to facilitate timely public health recognition and response to foodborne disease outbreaks and emerging foodborne diseases. Foodborne Pathog Dis. 2006;3(1):36–50.
7. Maiden M, Bygraves J, Feil EJ, Morelli G, Russell J, Urwin R, et al. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci USA. 1998;95(6):3140–5.
8. Spratt BG. Multilocus sequence typing: molecular typing of bacterial pathogens in an era of rapid DNA sequencing and the internet. Curr Opin Microbiol. 1999;2(3):312–6.
9. Maiden MC.. Multilocus sequence typing of bacteria. Annu Rev Microbiol. 2006;60:561–88.
10. Allerberger F. Molecular typing in public health laboratories: from an academic indulgence to an infection control imperative. J Prev Med Public Health. 2012;45(1):1–7.
11. Goering R, Morrison D, Doori Al Z, Edwards G, Gemmell C. Usefulness of mec-associated direct repeat unit (dru) typing in the epidemiological analysis of highly clonal methicillin-resistant Staphylococcus aureus in Scotland. Clin Microbiol Infect. 2008;14(10):964–9.
12. Frénay HM, Bunschoten AE, Schouls LM, van Leeuwen WJ, Vandenbroucke-Grauls CM, et al.. Molecular typing of methicillin-resistant Staphylococcus aureus on the basis of protein A gene polymorphism. Eur J Clin Microbiol Infect Dis. 1996;15(1):60–4.
13. Harmsen D, Claus H, Witte W, Rothganger J, Claus H, Turnwald D et al. Typing of methicillin-resistant Staphylococcus aureus in a university hospital setting by using novel software for spa repeat determination and database management. J Clin Microbiol. 2003;41(12):5442–8.
14. Oliveira DC, Santos M, Milheiriço C, Carriço JA, Vinga S, Oliveira AL, et al.. CcrB typing tool: an online resource for staphylococci ccrB sequence typing. J Antimicrob Chemother 2008;61(4):959–60.
15. Schouls LM, Spalburg EC, van Luit M, Huijsdens XW, Pluister GN, van Santen-Verheuvel MG, et al.. Multiple-locus variable number tandem repeat analysis of Staphylococcus aureus: comparison with pulsed-field gel electrophoresis and spa-typing. PLoS ONE 2009;4(4):e5082.
16. Grissa I, Bouchon P, Pourcel C, Vergnaud G. On-line resources for bacterial micro-evolution studies using MLVA or CRISPR typing. Biochimie. 2008;90(4):660–8.
17. Guigon G, Cheval J, Cahuzac R, Brisse S. MLVA-NET--a standardised web database for bacterial genotyping and surveillance. Euro Surveill. 2008;13(19) pii=18863. Available from: http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=18863
18. Allix-Béguec C, Harmsen D, Weniger T, Supply P, Niemann S. Evaluation and strategy for use of MIRU-VNTRplus, a multifunctional database for online analysis of genotyping data and phylogenetic identification of *Mycobacterium tuberculosis* complex isolates. J Clin Microbiol. 2008;46(8):2692–9.
19. Weniger T, Krawczyk J, Supply P, Niemann S, Harmsen D. MIRU-VNTRplus: a web tool for polyphasic genotyping of *Mycobacterium tuberculosis* complex bacteria. Nucleic Acids Res. 2010;38(Web Server issue):W326–31.
20. Jolley KA, Maiden MC. BIGSdb: Scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics. 2010;11:595.
21. Jolley KA, Bliss CM, Bennett JS, Bratcher HB, Brehony C, Colles FM, et al. Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. Microbiology 2012;158(pt 4):1005–15.
22. Sneath PHA, Sokal RR. Numerical Taxonomy: The principles and practice of numerical classification. 1st ed. San Francisco:W.H. Freeman & Co;1973.
23. Williams JG, Kubelik AR, Livak KJ, Rafalski JA, Tingey SV. DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic Acids Res. 1990;18(22):6531–5.
24. Vos P, Hogers R, Bleeker M, Reijans M, van De Lee T, Hornes M, et al. 1995. AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res. 1995;23(21):4407–14.
25. Feil EJ, Li BC, Aanensen DM, Hanage WP, Spratt BG. eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. J Bacteriol. 2004;186(5):1518–30.
26. Feil EJ, Holmes EC, Bessen DE, Chan MS, Day NP, Enright MC, et al. Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. Proc Natl Acad Sci U S A. 2001;98(1):182–7.
27. Smith JM, Feil EJ, Smith NH. Population structure and evolutionary dynamics of pathogenic bacteria. Bioessays. 2000;22(12):1115–22.
28. Francisco AP, Bugalho M, Ramirez M, Carriço JA. Global optimal eBURST analysis of multilocus typing data using a graphic matroid approach. BMC Bioinformatics. 2009;10:152.
29. Nübel U, Roumagnac P, Feldkamp M, Song J, Ko K, Huang Y, et al. Frequent emergence and limited geographic dispersal of methicillin-resistant Staphylococcus aureus. Proc Natl Acad Sci U S A. 2008;105(37):14130-5.
30. Roumagnac P, Weill F-X, Dolecek C, Baker S, Brisse S, Chinh NT, et al. Evolutionary history of Salmonella typhi. Science. 2006;314(5803):1301–4.
31. Baker S, Holt K, van de Vosse E, Roumagnac P, Whitehead S, King E, et al. High-throughput genotyping of Salmonella enterica serovar Typhi allowing geographical assignment of haplotypes and pathotypes within an urban District of Jakarta, Indonesia. J Clin Microbiol. 2008;46(5):1741–6.
32. Turner KM, Hanage WP, Fraser C, Connor TR, Spratt BG. Assessing the reliability of eBURST using simulated populations with known ancestry. BMC Microbiol. 2007;7:30.
33. McMillan DJ, Bessen DE, Pinho M, Ford C, Hall GS, Melo-Cristino J, et al. Population genetics of Streptococcus dysgalactiae subspecies equisimilis reveals widely dispersed clones and extensive recombination. PLoS One. 2010;5(7):e11741.
34. Martin DP, Lemey P, Lott M, Moulton V, Posada D, Lefeuvre P. RDP3: a flexible and fast computer program for analyzing recombination. Bioinformatics. 2010;26(19):2462–3.
35. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol. 2011;28(10):2731–9.
36. Sammeth M, Stoye J. Comparing tandem repeats with duplications and excisions of variable degree. IEEE/ACM Trans Comput Biol Bioinform. 2006;3(4):395–407.
37. Mellmann A, Weniger T, Berssenbrügge C, Rothgänger J, Sammeth M, Stoye J, et al. Based Upon Repeat Pattern (BURP): an algorithm to characterize the long-term evolution of Staphylococcus aureus populations based on spa polymorphisms. BMC Microbiol. 2007; 7:98.
38. Francisco AP, Vaz C, Monteiro PT, Melo-Cristino J, Ramirez M, Carriço JA. PHYLOViZ: Phylogenetic Inference and Data Visualization for Sequence Based Typing Methods. BMC Bioinformatics. 2012;13:87.
39. Grundmann H, Klugman KP, Walsh T, Ramon-Pardo P, Sigauque B, Khan W, et al. A framework for global surveillance of antibiotic resistance. Drug Resis Updat. 2011;14(2):79–87.
40. Grundmann H, Aanensen DM, van den Wijngaard CC, Spratt BG, Harmsen D, Friedrich AW, et al. Geographic distribution of Staphylococcus aureus causing invasive infections in Europe: a molecular-epidemiological analysis. PLoS Med. 2010;7(1):e1000215.
41. Aanensen DM, Huntley DM, Feil EJ, al-Own F, Spratt BG. EpiCollect: linking smartphones to web applications for epidemiology, ecology and community data collection. PLoS One. 2009;4(9):e6968.
42. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2009;155(2):945–59.
43. Didelot X, Falush D. Inference of bacterial microevolution using multilocus sequence data. Genetics. 2006;175(3):1251–66.

44. Corander J, Marttinen P. Bayesian identification of admixture events using multilocus molecular markers. Mol Ecol. 2006;15(10):2833–43.
45. Corander J, Tang J. Bayesian analysis of population structure based on linked molecular information. Math Biosci. 2007;205(1):19–31.
46. Tang J, Hanage WP, Fraser C, Corander J, Bourne PE. Identifying currents in the gene pool for bacterial populations using an integrative approach. PLoS Comput Biol. 2009; 5(8):e1000455.
47. Dale J, Price EP, Hornstra H, Busch JD, Mayo M, Godoy D, et al. Epidemiological tracking and population assignment of the non-clonal bacterium, Burkholderia pseudomallei. PLoS Negl Trop Dis. 2011;5(12):e1381.
48. Hanage WP, Fraser C, Tang J, Connor TR, Corander J. Hyper-recombination, diversity, and antibiotic resistance in pneumococcus. Science. 2009;324(5933):1454–7.
49. Wirth T, Hildebrand F, Allix-Béguec C, Wölbeling F, Kubica T, Kremer K, et al. Origin, spread and demography of the *Mycobacterium tuberculosis* complex. PLoS Pathog. 2008;4(9):e1000160.
50. Simpson E. Measurement of diversity. Nature. 1949;163:688.
51. Grundmann H, Hori S, Tanner G. Determining confidence intervals when measuring genetic diversity and the discriminatory abilities of typing methods for microorganisms. J Clin Microbiol. 2001;39(11):4190–2.
52. Rand WM. Objective criteria for the evaluation of clustering methods. Journal Am Statist Assoc.1971;66:846–50.
53. Hubert L, Arabie P. Comparing partitions. J Classif. 1985;2:193–218.
54. Wallace DL. A method for comparing two hierarchical clusterings: comment. J Am Statist Ass. 1983;78:569–76.
55. Carrico J, Pinto F, Simas C, Nunes S, Sousa N, Frazao N, et al. Assessment of band-based similarity coefficients for automatic type and subtype classification of microbial isolates analyzed by pulsed-field gel electrophoresis. J Clin Microbiol. 2005;43(11):5483–90.
56. Pinto FR, Melo-Cristino J, Ramirez M. A confidence interval for the wallace coefficient of concordance and its application to microbial typing methods. PLoS One. 2008;3(11):e3696.
57. Severiano A, Carriço JA, Robinson DA, Ramirez M, Pinto FR. Evaluation of jackknife and bootstrap for defining confidence intervals for pairwise agreement measures. PLoS One. 2011;6(5):e19539.
58. Severiano A, Pinto FR, Ramirez M, Carriço JA. Adjusted Wallace coefficient as a measure of congruence between typing methods. J Clin Microbiol. 2011;49(11):3997–4000.
59. Sabat AJ, Chlebowicz MA, Grundmann H, Arends JP, Kampinga G, Meessen NE, et al. Microfluidic-chip-based multiple-locus variable-number tandem-repeat fingerprinting with new primer sets for methicillin-resistant Staphylococcus aureus. J Clin Microbiol. 2012;50(7):2255–62.
60. Faria NA, Carriço JA, Oliveira DC, Ramirez M, de Lencastre H. Analysis of typing methods for epidemiological surveillance of both methicillin-resistant and methicillin-susceptible Staphylococcus aureus strains. J Clin Microbiol. 2008;46(1):136–144.
61. Miragaia M, Carrico JA, Thomas JC, Couto I, Enright MC, de Lencastre H. Comparison of Molecular Typing Methods for Characterization of *Staphylococcus epidermidis*: Proposal for Clone Definition. J Clin Microbiol. 2008;46(1):118–29.
62. Berners-Lee T, Hendler J. Publishing on the semantic web. Nature. 2001;410(6832):1023–4.
63. Bizer C, Heath T, Berners-Lee T. Linked data-the story so far. Int J Semantic Web and Inf Systems. 2009; 5(3).
64. Almeida JS, Chen C, Gorlitsky R, Stanislaus R, Aires-de-Sousa M, Eleutério P, et al. Data integration gets 'Sloppy'. Nat Biotechnol. 2006;24(9):1070–1.
65. Bard JB, Rhee SY. Ontologies in biology: design, applications and future challenges. Nat Rev Genet. 2004;5(3):213–22.
66. Almeida J, Tiple J, Ramirez M, Melo-Cristino J, Vaz C, P Francisco A, Carriço JA. An Ontology and a REST API for equence Based Microbial Typing Data, pp. 21–28. In: Freitas, A, Navarro, A, editors. Bioinformatics for Personalized Medicine. Berlin/Heidelberg:Springer,2012. .
67. Loman NJ, Constantinidou C, Chan JZ, Halachev M, Sergeant M, Penn CW, et al. High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. Nat Rev Microbiol. 2012;10(9):599–606.
68. Dunne WM, Westblade LF, Ford B. Next-generation and whole-genome sequencing in the diagnostic clinical microbiology laboratory. Eur J Clin Microbiol Infect Dis. 2012;31(8):1719–26.
69. Köser CU, Ellington MJ, Cartwright EJ, Gillespie SH, Brown NM, Farrington M, et al. Routine Use of Microbial Whole Genome Sequencing in Diagnostic and Public Health Microbiology. PLoS Pathog. 2012;8(8):e1002824.
70. Rohde H, Qin J, Cui Y, Li D, Loman NJ, Hentschke M, et al. Open-source genomic analysis of Shiga-toxin-producing E. coli O104:H4. N Engl J Med. 2011;365(8):718–24.
71. Mellmann A, Harmsen D, Cummings CA, Zentz EB, Leopold SR, Rico A, et al. Prospective genomic characterization of the German enterohemorrhagic Escherichia coli O104:H4 outbreak by rapid next generation sequencing technology. PLoS One. 2011;6(7):e22751.
72. Rasko DA, Webster DR, Sahl JW, Bashir A, Boisen N, Scheutz F, et al. Origins of the E. coli strain causing an outbreak of hemolytic-uremic syndrome in Germany. N Engl J Med. 2011;365(8):709–17.
73. Eyre DW, Golubchik T, Gordon NC, Bowden R, Piazza P, Batty EM, et al. A pilot study of rapid benchtop sequencing of Staphylococcus aureus and Clostridium difficile for outbreak detection and surveillance. BMJ Open. 2012;2(3).
74. Köser CU, Holden MT, Ellington MJ, Cartwright EJ, Brown NM, Ogilvy-Stuart AL, et al. Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. N Engl J Med. 2012;366(24):2267–75.
75. Harris SR, Feil EJ, Holden MT, Quail MA, Nickerson EK, Chantratita N, et al. Evolution of MRSA during hospital transmission and intercontinental spread. Science. 2012;327(5964):469–74.
76. Croucher N, Harris S, Fraser C, Quail M. Rapid pneumococcal evolution in response to clinical interventions. Science. 2011;331(6016):430–4.
77. Truman RW, Singh P, Sharma R, Busso P, Rougemont J, Paniz-Mondolfi A, et al. Probable zoonotic leprosy in the southern United States. N Engl J Med.2011;364(17):1626–33.
78. National Food Institute, DTU. Perspectives of a global, real-time microbiological genomic identification system. Brussels: DTU;2011.
79. Palm D, Johansson K, Ozin A, Friedrich AW, Grundmann H, Larsson JT, Struelens MJ. Molecular epidemiology of human pathogens: how to translate breakthroughs into public health practice, Stockholm, November 2011. Euro Surveill. 2012;17(2):pii=20054. Available from: http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=20054
80. Baker S, Hanage WP, Holt KE. Navigating the future of bacterial molecular epidemiology. Curr Opin Microbiol. 2010;13(5):640-5.